Separable Factor Analysis with Applications to Mortality Data

Bailey K. Fosdick[1] and Peter D. Hoff[1,2]

Departments of Statistics[1] and Biostatistics[2]

University of Washington

November 19, 2012

Abstract

Human mortality datasets can be expressed as multiway data arrays, the dimensions of which correspond to categories by which mortality rates are reported, such as age, sex, country and year. Regression models for such data typically assume an independent error distribution, or an error model that allows for dependence along at most one or two dimensions of the data array. However, failing to account for other dependencies can lead to inefficient estimates of regression parameters, inaccurate standard errors and poor predictions. An alternative to assuming independent errors is to allow for dependence along each dimension of the array using a separable covariance model. However, the number of parameters in this model increases rapidly with the dimensions of the array, and for many arrays, maximum likelihood estimates of the covariance parameters do not exist. In this paper, we propose a submodel of the separable covariance model that estimates the covariance matrix for each dimension as having factor analytic structure. This model can be viewed as an extension of factor analysis to array-valued data, as it uses a factor model to estimate the covariance along each dimension of the array. We discuss properties of this model as they relate to ordinary factor analysis, describe maximum likelihood and Bayesian estimation methods, and provide a likelihood ratio testing procedure for selecting the factor model ranks. We apply this methodology to the analysis of data from the Human Mortality Database, and show in a cross-validation experiment how it outperforms simpler methods. Additionally, we use this model to impute mortality rates for countries that have no mortality data for several years. Unlike other approaches, our methodology is able to estimate similarities between the mortality rates of countries, time periods, and sexes and use this information to assist with the imputations.

Keywords: array normal; Kronecker product; multiway data; Bayesian estimation; imputation.

This work was partially supported by NICHD grant R01HD067509.

1 Introduction

Human mortality data are used extensively by researchers and policy makers to analyze historic and current population trends and assess long-term impacts of public policy initiatives. To enable such inference, numerous regression models have been proposed that estimate mortality rates as a function of age using a small number of parameters (Heligman and Pollard (1980), Mode and Busby (1982), Siler (1983)). Practitioners using these methods typically model the age-specific death rates for each country, year, and sex combination separately and assume independent error distributions. Examples of death rates analyzed by such methods are shown in Figure 1 for the United States and Sweden. Each mortality curve is defined by 23 age-specific death rates and the average sex-specific mortality curve from 1960-1980 over thirty-eight countries is also displayed.

From the figure, it is clear that a country's mortality rates in one time period are similar to its rates in adjacent time periods. Acknowledging this fact, several researchers have developed models for "dynamic life tables", i.e. matrices of mortality rates for combinations of ages and time periods, for single country-sex combinations. An example of such a life table is the male death rates in Sweden from 1960 to 1980 shown in Figure 1. Some of the models developed for these data specify ARIMA processes for the time-varying model parameters (McNown and Rogers (1989), Renshaw and Haberman (2003b)), while others smooth the death rates over age and time using a kernel smoother (Felipe et al. (2001)), p-splines (Currie et al. (2004)), nonseparable age-time period covariance functions (Martínez-Ruiz et al. (2010)), or multiplicative effects for age and time (Lee and Carter (1992), Renshaw et al. (1996), Renshaw and Haberman (2003a), Renshaw and Haberman (2003c), Chiou and Müller (2009)).

Human mortality datasets typically provide mortality rates of populations corresponding to combinations of several factors. For example, the Human Mortality Database (HMD, Human Mortality Database, 2011) provides mortality rates of populations corresponding to combinations of 40 countries, 9 time periods, 23 age groups, and both male and female sexes. As is shown in Figure 1, mortality rates of men and women within a country will typically both be higher than or both lower than the sex-specific rates averaged across countries. Furthermore, differences between male and female mortality rates generally show trends that are consistent across countries and time periods. Such patterns suggest joint estimation of mortality rates using a model that can share information across levels of two or more factors. Two models that consider death rates for more

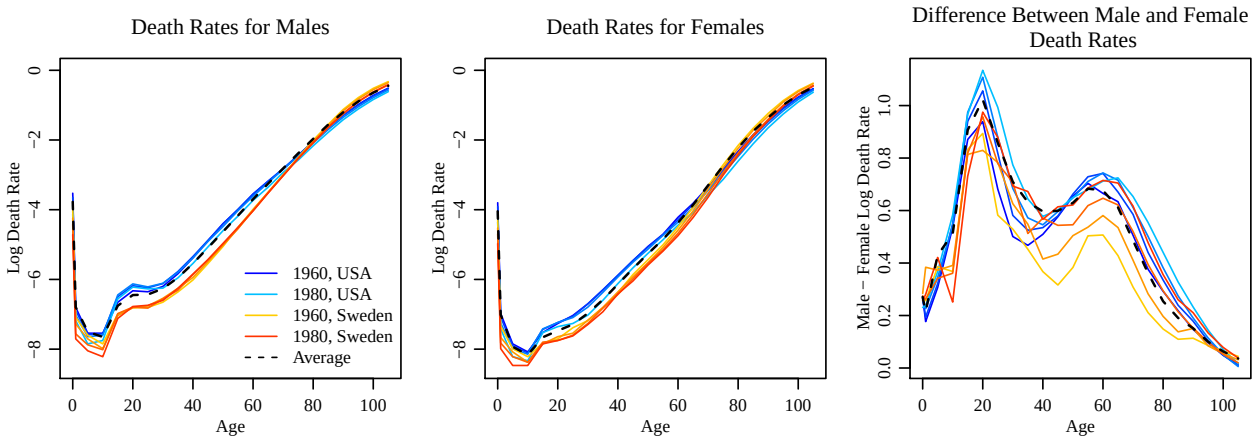

Figure 1: Mortality curves for the United States of America and Sweden. The gradient of colors for each country represents the log death rates in the four 5-year time periods from 1960 to 1980. The average sex-specific mortality curve over the four time periods and all countries is shown in black.

than one country or sex are that developed by Li and Lee (2005), which estimates common age and time period effects for a group of countries or both sexes, and Carter and Lee (1992), where male and female death rates within the same country share a time-varying mortality level. Although these methods consider either both sexes or multiple countries, the extreme similarity of the curves in Figure 1 for males across countries and for a given country across sexes suggest that separately modeling death rates for different countries or sexes is inefficient, and inference may be improved by using a joint model that shares information across all factors.

With this in mind, we consider a regression model for the HMD data consisting of a mean model that is a piecewise-polynomial in age with additive effects for country, time period, and sex (more details on this model, and its comparison to other models, are provided in Section 4). This mean model is extremely flexible: It contains over 370 parameters and an ordinary least squares (OLS) fit accounts for over 99% of the total variation in the data (coefficient of determination, $R^2 > 0.995$). Nonetheless, an analysis of the residuals from the OLS fit indicates that some clear patterns in the data are not captured by the regression model, and in particular, a model of independent errors is a poor representation of these data. To illustrate this, note that the residuals can be represented as a 4-way array, the dimensions of which are given by the number of levels of each of the four factors: country, time period, sex and age. To examine residual correlation across levels of a factor,

the 4-way array of residuals can be converted into a matrix whose columns represent the levels of the factor, and a sample correlation matrix for the factor can be obtained. Figure 2 summarizes the patterns in the residual correlations using the first two principal components of each sample correlation matrix. If a model of independent errors were to be adequate, we would expect the sample correlation values to be small and centered about zero, and no discernible patterns to exist in the principal components. However, the sample correlations are substantially more positive than would be expected under independence: 59% of the observed country correlations, 61% of time period correlations and 98% of age correlations are greater than the corresponding 95% theoretical percentiles under the independence assumption. Additionally, there are clear geographic, temporal, and age trends in the principal components in Figure 2. For example, the residuals for the Ukrainian mortality rates are positively correlated to those for Russia, and the residuals for the year 2000 are positively correlated with those for 1995. This residual analysis suggests that an assumption of uncorrelated errors is inappropriate on account of the positive correlation exhibited by residuals across levels of the factors.

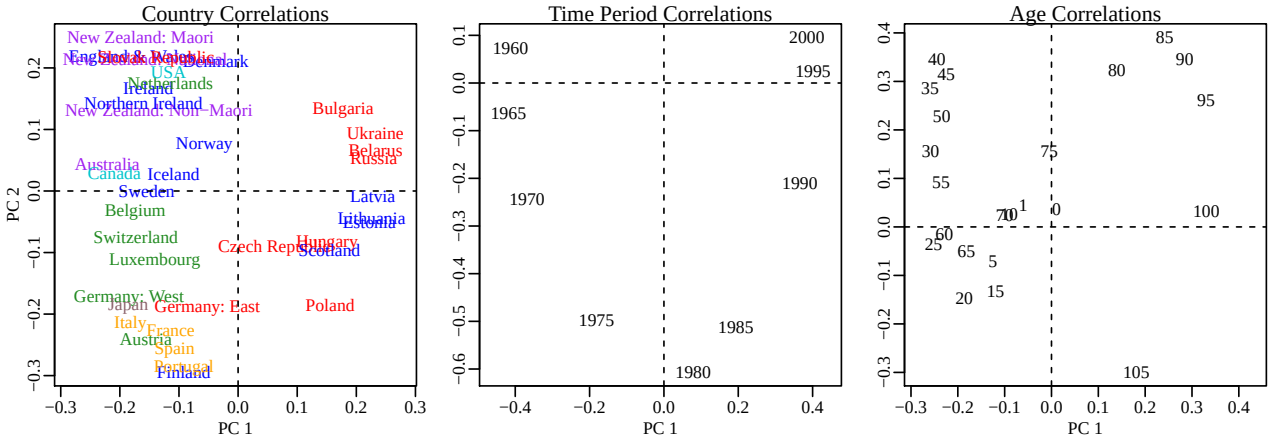

Figure 2: The first two principal components of each sample correlation matrix are displayed, and countries in the same United Nations region are shown in the same color. Close proximity in the principal components space away from the origin is indicative of a positive correlation.

Failure to recognize correlated errors can lead to a variety of inferential problems, such as inefficient parameter estimates and inaccurate standard errors. For the analysis of the mortality data, an additional important consequence is that the accuracy of predictions of missing mortality rates may suffer. Predicting missing death rates is a primary application of modeling mortality

data, as developing countries often lack reliable death registration data. It is possible that the residual dependence could be reduced by increasing the flexibility of the mean model, but since this is already fairly complex, we may instead prefer to represent residual dependence with a covariance model, leading to a general linear model for the data in which the mean function and residual covariance are estimated simultaneously.

The mortality data, like the residuals, can be represented as a 4-way array, each dimension of which corresponds to one of the factors of country, time period, sex and age. In the literature on multiway array data (see, for example, Kroonenberg (2008)), each dimension is referred to as a mode of the array, so a 4-way array of mortality data consists of four modes. As described by Hoff (2011), a natural covariance model for a K-way data array is a separable covariance model, parameterized in terms of K covariance matrices, one for each mode of the array. If the array is also assumed to be normally distributed, the model is referred to as the array normal model, and can be seen as an extension of the matrix normal model (Dawid (1981)).

Even though the separable covariance model is not a full, unstructured covariance model, the array normal likelihood is unbounded for many array dimensions, prohibiting the use of maximum likelihood methods (Manceur and Dutilleul (2013)). Estimates of the array normal covariance parameters can still be obtained by taking a Bayesian approach (Hoff (2011)) or by using a penalized likelihood (Allen and Tibshirani (2010)). However, the lack of existence of the maximum likelihood estimates (MLEs) indicates that the data is unable to provide information about all of the parameters. In this article we propose an alternative modeling approach that parameterizes the covariance matrix of each mode as a reduced rank matrix plus a diagonal matrix. This covariance structure is commonly associated with factor analysis, and is referred to here as factor analytic covariance structure. We call this new model the Separable Factor Analysis (SFA) model, as it is an extension of factor analysis to array-valued data. The reduction in the number of parameters by using covariance matrices with factor analytic structure leads to existence of MLEs for the SFA parameters in many cases when the MLEs of the array normal parameters do not exist, as well as a parsimonious representation of mode-specific covariance in an array-valued dataset.

This article is outlined as follows: In the next section we introduce and motivate SFA, as well as discuss its properties and similarities to ordinary factor analysis. We describe two estimation procedures in Section 3: an iterative maximum likelihood algorithm and a Metropolis-Hastings

sampler for inference in a Bayesian framework. A likelihood ratio testing procedure for selecting the rank of the factor model for each mode is also presented. In Section 4 the SFA model is used to analyze the HMD mortality data and its performance is compared to simpler covariance models in a simulation study. We illustrate how SFA uses estimated similarities between country mortality rates to provide imputations for countries missing mortality data for several years. This prediction method extends the approach taken in Coale and Demeny (1966), Brass (1971), United Nations (1982), and Murray et al. (2003), where one country's mortality curve is modeled a function of another's. Our approach is novel in that it estimates the covariance between mortality rates across all countries, time periods, and sexes, and uses these relationships to impute missing death rates. We conclude with a discussion in Section 5.

2 Extending factor analysis to arrays

2.1 Motivating separable factor analysis

Suppose Y is a K-way array of dimension $m_1 \times m_2 \times ... \times m_K$. We are interested in relating the data Y to explanatory variables X through the model $Y = M(X, \beta) + E$, where β represents unknown regression coefficients and E represents the deviations from the mean. As was discussed in the preliminary analysis of the mortality data in Section 1, it is often unreasonable to assume the elements of E are independent and identically distributed.

In cases where there is no independent replication, estimation of the $\text{Cov}[E]$ can be problematic as it must be based on essentially a single sample. One solution is to approximate the covariance matrix with one with simplified structure. A frequently used model in spatio-temporal analysis is a separable covariance model (Stein (2005), Genton (2007)). This model estimates a covariance matrix for each mode of the array and is written $\text{Cov}[\text{vec}(E)] = \Sigma_K \otimes \Sigma_{K-1} \otimes ... \otimes \Sigma_1$, where "vec" and "$\otimes$" denote the vectorization and Kronecker operators, respectively. In the context of the mortality data, this model contains a covariance matrix for country (Σ_c), time period (Σ_t), age (Σ_a), and sex (Σ_s). A separable covariance model with the assumption that the deviations are normally distributed, $\text{vec}(E) \sim \text{normal}(0, \text{Cov}[\text{vec}(E)])$, is an array normal model and was developed by Hoff (2011) as an extension of the matrix normal (Dawid (1981), Browne (1984), Oort (1999)).

The mode covariance matrices in the array normal model are not estimable for certain array dimensions using standard techniques such as maximum likelihood estimation (Manceur and Dutilleul (2013)). However, often the covariance matrices of large modes can be well approximated by matrices with simpler structure. A common approach in the social sciences to modeling the covariance matrix of a high-dimensional random vector is to use factor analysis. The standard k-factor model for a random vector $x \in \mathbb{R}^p$ parameterizes the covariance matrix as $\text{Cov}[x] = \Lambda\Lambda^T + D^2$, where $\Lambda \in \mathbb{R}^{p\times k}$, $k < p$, and D is a diagonal matrix (Spearman (1904), Mardia et al. (1979)). We will refer to this model as single mode factor analysis as it models the covariance among one set of variables. When the number of independent observations n is less than p, the sample covariance matrix is not positive definite and hence cannot be used as an estimate of $\text{Cov}[x]$. Nevertheless, under the assumption that x follows a multivariate normal distribution with known mean, the maximum likelihood estimate of the factor analytic covariance matrix exists if $k < \min(p, n)$ (Robertson and Symons (2007)).

We propose a submodel of the array normal model where each mode covariance matrix potentially has factor analytic structure. We call this model *Separable Factor Analysis (SFA)* and it is written as follows:

$$\begin{aligned} \text{vec}(E) &\sim \text{normal}(0, \text{Cov}[\text{vec}(E)]) \\ \text{Cov}[\text{vec}(E)] &= \Sigma_K \otimes \Sigma_{K-1} \otimes ... \otimes \Sigma_1, \\ \text{where } \Sigma_i &= \Lambda_i\Lambda_i^T + D_i^2 \ \text{ for } 0 \le k_i < m_i \end{aligned} \tag{1}$$

and Σ_i is unconstrained (i.e. equals any positive definite matrix) if $k_i = m_i$. SFA models are characterized by the covariance matrix structure chosen for each mode and can be represented by a K-vector of ranks $(k_1, ..., k_K)$, where k_i equals the rank of Λ_i if mode i's covariance matrix has factor analytic structure and equals m_i if the mode covariance matrix is unstructured. Note that we consider the $k_i = 0$ case where the covariance matrix is diagonal, reflecting independence of entries along the mode. A k_i-factor analytic covariance matrix, $\Lambda_i\Lambda_i^T + D_i^2$, contains $\delta(m_i, k_i) = \left[(m_i - k_i)^2 - (m_i + k_i)\right]/2$ fewer parameters than an unstructured $m_i \times m_i$ covariance matrix. If $\delta(m_i, k_i) \le 0$, the factor analytic covariance matrix does not provide reduced structure, and to prevent overparameterizing the model, either the factor analytic rank, k_i, should be decreased or an unstructured covariance matrix should be specified.

SFA has advantages over the array normal model that stem from it being an extension of factor analysis to multiple modes. Each mode covariance matrix can be interpreted independently as a decomposition of the mode variability into shared and unique latent components as in classical factor analysis (see Properties below). In addition, empirical evidence has shown that MLEs of the SFA covariance parameters exist for array dimensions where the MLEs of the array normal unstructured covariance matrices do not exist.

2.2 Properties of SFA

In this section we relate the SFA parameters to those in ordinary factor analysis, discuss indeterminancies in the model, and interpret the SFA parameters when the true covariance matrix in each mode is unstructured. We use the concept of array matricization to describe how these properties relate to each mode. Here we define matricizing an array in the ith mode as unfolding the array into a matrix $Y_{(i)}$ of dimension $(m_i \times \prod_{j \neq i} m_j)$, where an earlier mode index always varies faster than a later mode index in the columns (Kolda and Bader (2009)). There are multiple ways to matricize an array but here we follow this convention (Kiers (2000), De Lathauwer et al. (2000)).

Latent variable representation

A single mode k-factor model for a sample of n mean zero p-variate random vectors is written $\{x_1, ..., x_n\} \sim$ i.i.d. normal$(0, \Lambda\Lambda^T + D^2)$, where $\Lambda \in \mathbb{R}^{p \times k}$ and D is a diagonal matrix. Collecting the random vectors in a $p \times n$ matrix $X = [x_1, ..., x_n]$, this model has an equivalent latent variable representation as a decomposition into common latent factors, $Z = [z_1, ..., z_n]$, and variable specific latent factors, $E = [e_1, ..., e_n]$, as follows.

$$
\begin{aligned}
& X_{p\times n} = \Lambda_{p\times k} Z_{k\times n} + D_{p\times p} E_{p\times n} \\
& \{z_1, ..., z_n\} \sim \text{i.i.d. normal}(0, \mathrm{I}_k) \qquad \mathrm{Cov}[z_i, e_j] = 0_{k\times p} \qquad \text{for all } i, j \\
& \{e_1, ..., e_n\} \sim \text{i.i.d. normal}(0, \mathrm{I}_p)
\end{aligned} \tag{2}
$$

In this representation the jth observation of the ith variable X_{ij} is written as a linear combination of common latent factors z_i with coefficients given by the ith row of Λ, plus a single variable specific factor E_{ij}, whose coefficient is the ith diagonal element of D.

A similar representation exists for each mode with a factor analytic covariance structure in the

SFA model. Consider a mean zero array Y and an SFA model with a factor analytic covariance matrix in the ith mode. Define $\widetilde{Y}^i$ to be the array obtained by standardizing Y with all but the ith mode's covariance matrix:

$$\text{vec}(\widetilde{Y}^i) \coloneqq \text{vec}(Y)(\Sigma_K^{-1/2} \otimes ... \otimes \Sigma_{i+1}^{-1/2} \otimes \mathrm{I}_{m_i} \otimes \Sigma_{i-1}^{-1/2} \otimes ... \otimes \Sigma_1^{-1/2}). \tag{3}$$

It follows that

$$\widetilde{Y}^i_{(i)} = [y_1, ..., y_{m_{-i}}] \overset{d}{=} \Lambda_i Z^i + D_i E^i \quad \text{and} \quad \{y_1, ..., y_{m_{-i}}\} \sim \text{i.i.d. normal}(0, \Lambda_i \Lambda_i^T + D_i^2) \tag{4}$$

where $m_{-i} = \prod_{j \neq i} m_j$, and Z^i and E^i are $k_i \times m_{-i}$ and $m_i \times m_{-i}$, respectively, with the same distributional properties as Z and E in (2). The superscript i on $\widetilde{Y}^i_{(i)}$ indicates the ith mode has not been standardized and the subscript (i) indicates the array has been matricized along the ith mode. The representation in (4) suggests the parameters $\{\Lambda_i, D_i\}$ can be viewed as single mode factor analysis parameters for the ith mode of the array when the covariance in all other modes has been removed.

Model indeterminacies

SFA as parameterized in (1) has two indeterminacies, one of which is common to all factor models and one that is common to all array normal models. The first indeterminacy, which is also present in single mode factor analysis, is the orientation of the Λ matrices. The array covariance matrix in (1) is the same with mode i factor analytic parameters $\{\Lambda_i, D_i\}$ as it is with parameters $\{\Lambda_i G_i, D_i\}$, where G_i is any $k_i \times k_i$ orthogonal matrix. The second indeterminacy concerns the scales of the mode covariance matrices and stems from the model's separable covariance structure. The scale of a mode's covariance matrix can be moved to another mode covariance matrix or split among multiple modes' covariance matrices without changing the model. For example, the transformation $\{\Sigma_i, \Sigma_j\} \mapsto \{c\Sigma_i, \Sigma_j/c\}$ does not affect the array covariance matrix for any $c > 0$. This scale non-identifability is eliminated if all mode covariance matrices are restricted to have trace equal to one and a scale parameter is included for the total variance of the array.

Pseudo-true parameters

In single mode factor analysis the goal is to represent the covariance among a large set of variables

in terms of a small number of latent factors. However, often it is unlikely the true covariance matrix has factor analytic structure. Consider a $p \times n$ matrix $X = [x_1, ..., x_n]$ and suppose $\{x_1, ..., x_n\} \sim$ i.i.d. normal$(0, \Sigma)$. There is interest in what k-factor analytic parameter values, Λ and D, best approximate the true covariance matrix Σ. These optimal parameter values, denoted $\overline{\Lambda}(\Sigma)$ and $\overline{D}(\Sigma)$, are those that minimize the Kullback-Leibler (KL) divergence between the factor model and the multivariate normal model. Minimizing the KL divergence is equivalent to maximizing the expected value of the k-factor analysis (FA) probability density with respect to the true multivariate normal (MN) distribution. Thus, $\overline{\Lambda}(\Sigma)$ and $\overline{D}(\Sigma)$ can be defined as

$$\{\overline{\Lambda}(\Sigma), \overline{D}(\Sigma)\} := \underset{\Lambda, D}{\operatorname{argmax}} \, \mathrm{E}_{\mathrm{MN}}[p_{\mathrm{FA}}(X|\Lambda, D)] = \underset{\Lambda, D}{\operatorname{argmax}} \; c_{FA} - \frac{n}{2}\log(|\Lambda\Lambda^T + D^2|) - \frac{n}{2}\mathrm{tr}[(\Lambda\Lambda^T + D^2)^{-1}\Sigma]$$

where "tr" represents the trace operator and c_{FA} is a constant not depending on Λ or D. The diagonal matrix D that best approximates the true covariance matrix in the case of $k = 0$ is given by $\overline{D}(\Sigma) = \mathrm{diag}(\Sigma)^{1/2}$, where "diag" is the operator that replaces all off-diagonal entries with zero.

Similar to single mode factor analysis, SFA is an approximation to a separable covariance structure and modes' true covariance matrices are unlikely to have factor analytic structure. Suppose the distribution of Y is array normal with mean zero and covariance matrices $\widetilde{\boldsymbol{\Sigma}} = \{\widetilde{\Sigma}_i : 1 \leq i \leq K\}$. Consider a $(k_1, ..., k_K)$ SFA model for Y with parameters $\boldsymbol{\Lambda} = \{\Lambda_i : 0 < k_i < m_i\}$, $\boldsymbol{D} = \{D_i : 0 \leq k_i < m_i\}$, and $\boldsymbol{\Sigma} = \{\Sigma_j : k_j = m_j\}$. The expected value of the SFA probability density with respect to the true array normal (AN) model is

$$\mathrm{E}_{\mathrm{AN}}[p_{\mathrm{SFA}}(Y|\boldsymbol{\Sigma}, \boldsymbol{D}, \boldsymbol{\Lambda})] = c_{SFA} - \sum_{i=1}^{K} \frac{m}{2m_i}\log(|\Sigma_i|) - \frac{1}{2}\prod_{i=1}^{K}\mathrm{tr}[\Sigma_i^{-1}\widetilde{\Sigma}_i], \tag{5}$$

$$\text{where} \;\; \Sigma_i = \Lambda_i\Lambda_i^T + D_i^2 \;\; \text{for} \;\; 0 \leq k_i < m_i,$$

c_{SFA} is a constant independent of the SFA parameters, and $m = \prod_{i=1}^{K} m_i$. Let $\overline{\overline{\Lambda}}(\widetilde{\boldsymbol{\Sigma}})$, $\overline{\overline{D}}(\widetilde{\boldsymbol{\Sigma}})$, and $\overline{\overline{\Sigma}}(\widetilde{\boldsymbol{\Sigma}})$ denote the SFA parameters that maximize (5) and, hence, provide the best approximation to the true separable covariance matrix. It can be shown that for all appropriate i, j, and k

$$\overline{\overline{\Lambda}}_i(\widetilde{\boldsymbol{\Sigma}}) = \overline{\Lambda}(\widetilde{\Sigma}_i), \qquad \overline{\overline{D}}_j(\widetilde{\boldsymbol{\Sigma}}) = \overline{D}(\widetilde{\Sigma}_j), \qquad \text{and} \qquad \overline{\overline{\Sigma}}_k(\widetilde{\boldsymbol{\Sigma}}) = \widetilde{\Sigma}_k. \tag{6}$$

This means that the best factor analytic parameters for a given mode in the SFA model are the closest fitting single mode factor analytic parameters for that mode's true covariance matrix. Furthermore, as we might expect, the optimal values of the unstructured covariance matrices in the

SFA model are the modes' true covariance matrices. This implies that when the true model is array normal, the optimal SFA parameters for a given mode do not depend on the specified covariance structures in the other modes. Note that the scale indeterminacy of the covariance matrices is still present here. Thus, there is a set of optimal SFA parameter values that provide the same approximation and can be derived from one another by reallocating the covariance matrices' scales. Asymptotically, as the number of replicates of the array increases, these optimal SFA parameter values are the limiting values of the SFA maximum likelihood estimates (White (1982)).

3 Estimation and testing

In this section we consider parameter estimation for the SFA model and propose a likelihood ratio testing procedure for selecting the ranks $(k_1, ..., k_K)$. Two estimation methods are described here: an iterative algorithm for maximum likelihood estimation and a Metropolis-Hastings algorithm which approximates the posterior distribution of the parameters given the data. For notational convenience we present the case where the array has mean zero, however both estimation methods and the testing procedure can be extended to allow for a mean structure. Examples of such extensions are discussed in Section 4 for the mortality data.

3.1 Maximum likelihood estimation

Simultaneous maximization of the SFA log likelihood with respect to all parameters is difficult. However, the maximization steps are manageable if done separately for each mode's covariance parameters. We propose an iterative algorithm that at each step maximizes the SFA log likelihood over a single mode's covariance parameters using the latest values of all other modes' parameters. This algorithm can be viewed as a form of block coordinate ascent and is guaranteed to increase the log likelihood at each step.

Let $\mathbf{\Lambda} = \{\Lambda_i : 0 < k_i < m_i\}$, $\boldsymbol{D} = \{D_i : 0 \leq k_i < m_i\}$, and $\mathbf{\Sigma} = \{\Sigma_j : k_j = m_j\}$ as in Section 2.2. Also let $\mathbf{\Lambda}_{-\boldsymbol{j}} = \mathbf{\Lambda}/\{\Lambda_j\}$ be the set $\mathbf{\Lambda}$ with Λ_j removed, and define $\boldsymbol{D}_{-\boldsymbol{j}}$ and $\mathbf{\Sigma}_{-\boldsymbol{i}}$ analogously. The iterative maximum likelihood algorithm proceeds as follows.

0. Specify initial values for all covariance parameters $\{\mathbf{\Lambda}, \boldsymbol{D}, \mathbf{\Sigma}\}$.

1. For each mode $\{i : k_i = 0\}$, update the estimate of D_i.

2. For each mode $\{i : 0 < k_i < m_i\}$, update the estimates of Λ_i and D_i.

3. For each mode $\{i : k_i = m_i\}$, update the estimate of Σ_i.

4. Repeat steps 1-3 until a desired level of convergence is obtained.

The maximization in steps 1 and 3 over a diagonal matrix and an unstructured covariance matrix, respectively, are straightforward. The SFA log likelihood as a function of D_i, treating all other parameters as fixed, is written

$$\ell(D_i|\boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \boldsymbol{D_{-i}}, Y) = b_i - \frac{m}{2m_i}\log(|D_i^2|) - \frac{1}{2}\mathrm{tr}[\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T D_i^{-2}] \tag{7}$$

where b_i is a constant independent of D_i and $m = \prod_i^K m_i$. The maximizing value of D_i and update for step 1 is thus

$$D_i^2 = \mathrm{diag}\left(\frac{m_i}{m}\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T\right)$$

where the covariance matrices used to standardize Y in $\widetilde{Y}^i$ are the latest covariance matrix estimates. The SFA log likelihood as a function of an unstructured covariance matrix Σ_i is given by (7) where D_i^2 is replaced by Σ_i. Hence, the value of Σ_i that maximizes the corresponding log likelihood and the update for step 3 is

$$\Sigma_i = \frac{m_i}{m}\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T.$$

Estimation of a mode's factor analytic parameters in step 2 is more difficult, but can be accomplished using methods developed for single mode factor analysis. The SFA log likelihood as a function of the ith mode's factor analytic parameters, treating all other modes' parameters as fixed, is

$$\ell(\Lambda_i, D_i|\boldsymbol{\Sigma}, \boldsymbol{\Lambda_{-i}}, \boldsymbol{D_{-i}}, Y) = c_i - \frac{m}{2m_i}\log(|\Lambda_i\Lambda_i^T + D_i^2|) - \frac{1}{2}\mathrm{tr}[(\Lambda_i\Lambda_i^T + D_i^2)^{-1}\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T] \tag{8}$$

where c_i is a constant not depending on Λ_i or D_i. The log likelihood for a single mode k_i-factor model for a $p \times n$ matrix X is written

$$\ell(\Lambda, D|X) = c - \frac{n}{2}\log(|\Lambda\Lambda^T + D^2|) - \frac{1}{2}\mathrm{tr}[(\Lambda\Lambda^T + D^2)^{-1}XX^T]. \tag{9}$$

Notice that the SFA log likelihood has the the same form as that for single mode factor analysis where XX^T and n are replaced by $\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T$ and m/m_i, respectively. Therefore, we can use existing estimation methods for single mode factor analysis to update Λ_i and D_i.

Numerous iterative algorithms have been developed to obtain the single mode factor model maximum likelihood estimates; however many suffer from poor convergence behavior (Lawley (1940), Jöreskog (1967), Jennrich and Bobinson (1969)). An expectation-maximization (EM) algorithm was developed based on the model representation in (2) that treats Z as latent variables (Dempster et al. (1977), Rubin and Thayer (1982)). The slow convergence of this algorithm led to expectation/conditional maximization either (ECME) algorithms, some of which rely on numerical optimization procedures (Liu and Rubin (1998), Zhao et al. (2008)). Zhao et al. (2008) proposed an iterative algorithm that updates Λ, treating D as known, and then sequentially updates each diagonal element of D, treating Λ and all other elements of D as known. This algorithm has closed form expressions for all parameter updates and was shown to outperform the EM algorithm and its extensions in terms of convergence and computation time. For these reasons, we chose to use it for step 2 of the SFA estimation procedure.

Divergence of the SFA maximum likelihood algorithm, where the log likelihood continually grows at a nondecreasing rate, is evidence that the maximum likelihood estimates do not exist. While the update in step 1 for a mode with a diagonal covariance matrix is alway well defined (i.e. the SFA log likelihood in (7) has a maximum in terms of D_i), step 2 of the algorithm for an unstructured covariance matrix is only well defined if $m_i < \prod_{j \neq i} m_j$. Similarly, step 3 is well defined for a mode i with a factor analytic covariance matrix if $k_i < \text{rank}(\widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T)$. This latter requirement is effectively equivalent to $k_i < \min(m_i, \prod_{j \neq i} m_j)$ since $\widetilde{Y}^i_{(i)}$ is unlikely to be rank deficient for a continuous array Y.

3.2 Bayesian estimation

Mortality information is limited for many undeveloped countries that do not have reliable death registration data. Thus, it is not uncommon to be missing a country's death rates for specific ages or at all ages in a given year. A Bayesian approach to parameter estimation can easily accommodate missing data and provide predictive distributions of the missing values. Let Y_o denote the portions of the array Y that are observed and Y_m represent those values that are missing. Inference for the parameters and the missing data can be based on the joint posterior distribution of the parameters and missing data given the observed data, $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}, Y_m | Y_o)$. This posterior distribution is written $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}, Y_m | Y_o) \propto p(Y | \boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}) p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma})$, where $p(Y | \boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma})$ is the density of the $(k_1, ..., k_K)$

SFA model for $Y = \{Y_o, Y_m\}$ and $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma})$ is the joint prior distribution of the parameters. In the case of no missing data, one can consider Y_m to be the empty set and Y_o to equal Y.

Prior specification

In the absence of real prior information, we suggest a convenience prior composed of semiconjugate distributions for the parameters, which also reflects some of the indeterminacies in the model. For an SFA model in which mode i's covariance matrix is unstructured, the prior distribution for Σ_i^{-1} is Wishart$(\kappa_i, \mathrm{I}_{m_i})$ with hyperparameter κ_i, where $\kappa_i \geq m_i$. For a mode i with a factor analytic covariance matrix, the joint prior distribution of $\{\Lambda_i, D_i\}$ is specified by the marginal distribution of D_i and the conditional distribution of Λ_i given D_i, as follows:

$$\{\mathrm{vec}(\Lambda_i)|D_i\} \sim \mathrm{normal}(0, \mathrm{I}_{k_i} \otimes D_i^2) \tag{10}$$

$$\{D_i^{-2}[1,1], ..., D_i^{-2}[m_i, m_i]\} \sim \text{i.i.d. gamma}(\nu_0/2,\ \text{rate} = \nu_0 d_0^2/2) \qquad \nu_0 > 0,\ d_0^2 > 0. \tag{11}$$

A priori each mode's parameters are modeled as independent of all other modes' parameters given the hyperparameters ν_0, d_0^2, and $\{\kappa_i : k_i = m_i\}$. Thus, the joint prior distribution $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma})$ is equal to the product of the marginal distributions of each mode's parameters.

The prior distribution of the factor analytic parameters given in (10) and (11) has nice properties related to the rotational indeterminacies in the $\boldsymbol{\Lambda}$ matrices. Recall that the SFA likelihood is invariant to rotation of Λ_i, meaning $\mathrm{L}_{\mathrm{SFA}}(\Lambda_i, D_i, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}_{-i}, \boldsymbol{D}_{-i}|Y) =$
$\mathrm{L}_{\mathrm{SFA}}(\Lambda_i G_i, D_i, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}_{-i}, \boldsymbol{D}_{-i}|Y)$ where $\mathrm{L}_{\mathrm{SFA}}$ is the SFA likelihood and G_i is any $k_i \times k_i$ orthogonal matrix. If the joint prior distribution $p(\Lambda_i, D_i)$ is integrated over the diagonal elements of D_i, the marginal distribution of Λ_i is obtained and can be expressed as

$$p(\Lambda_i) \propto \prod_{j=1}^{m_i} \left[\nu_0 d_0^2 + ||\Lambda_i[j,]||^2\right]^{(k_i+\nu_0)/2},$$

where $||\cdot||^2$ denotes the Frobenius norm. Observe that $p(\Lambda_i) = p(\Lambda_i G_i)$ implying that the prior distribution is also invariant to rotations of Λ_i. This is a desirable property as it indicates the prior does not favor one set of parameters over another if they are equivalent given the data (i.e. have the same SFA likelihood). Namely, all parameter values $\{\Lambda_i G_i : G_i^T G_i = G_i G_i^T = I\}$ are given equal probability in the prior.

Metropolis-Hastings algorithm

The posterior distribution $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}, Y_m|Y_o)$ is not a standard distribution and is difficult to sample from directly, so a Metropolis-Hastings algorithm is used to obtain a Monte Carlo approximation of it. The Metropolis-Hastings algorithm produces a Markov chain of $\{\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}, Y_m\}$, whose stationary distribution is equal to $p(\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}, Y_m|Y_o)$. The algorithm proceeds by iteratively proposing new values of the missing data and each mode's parameters, and accepting or rejecting the proposals based on an acceptance probability. The algorithm can be described as follows:

0. Specify initial values for all covariance parameters $\{\boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{\Sigma}\}$ and missing data Y_m.
1. For each mode $\{i : k_i = 0\}$, update D_i.
2. For each mode $\{i : 0 < k_i < m_i\}$, update Λ_i and D_i.
3. For each mode $\{i : k_i = m_i\}$, update Σ_i.
4. If elements of Y are missing, update Y_m.
5. Repeat steps 1-4 until a sufficiently accurate approximation of the posterior distribution is obtained.

The update of D_i in step 1 is straightforward since the full conditional distribution of the entires in D_i^{-2} given the data and all other parameters is a product of the following independent gamma distributions:

$$\{D_i^{-2}[j,j]|\boldsymbol{D_{-i}}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, Y\} \sim \text{gamma}\left((\nu_0 + m/m_i)/2, \text{rate} = \left(\nu_0 d_0^2 + S_i[j,j]\right)/2\right) \tag{12}$$

for $j \in \{1, ..., m_i\}$ where $S_i = \widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T$. The Metropolis-Hastings acceptance probability is equal to one when sampling from parameters' full conditional distribution. Thus, the update in step 1 is performed by setting the new value of D_i equal to a sample from (12), where Y is comprised of Y_o and the most current update of Y_m, and the covariance matrices used to standardize Y in $\widetilde{Y}^i$ are the most current parameter updates.

The updates for the factor analytic parameters $\{\Lambda_i, D_i\}$ in step 2 are based on the latent variable representation of SFA introduced in (4), which is that $\widetilde{Y}^i_{(i)} \stackrel{d}{=} \Lambda_i Z^i + D_i E^i$ where the elements of Z^i and E^i are independent standard normal random variables. Conditioning on Z^i, the mode i

k_i-factor model can be written

$$\{\text{vec}(\widetilde{Y}^i_{(i)})|Z^i, \Lambda_i, D_i\} \sim \text{normal}(\text{vec}(\Lambda_i Z^i), \text{I}_{m/m_i} \otimes D_i^2).$$

Using this representation of the model, we developed the following Metropolis-Hastings updates:

A. Update Λ_i as follows.

(i) Sample $\{\text{vec}(Z^i)|\Lambda_i, D_i, \widetilde{Y}^i\} \sim \text{normal}(\text{vec}(\phi\Lambda_i^T D_i^{-2}\widetilde{Y}^i_{(i)}), \text{I}_{m/m_i} \otimes \phi)$
where $\phi = (\Lambda_i^T D_i^{-2}\Lambda_i + \text{I})^{-1}$.

(ii) Sample

$$\{\text{vec}(\Lambda_i)|Z^i, Y, \boldsymbol{D}, \boldsymbol{\Lambda_{-i}}, \boldsymbol{\Sigma}\} \sim \text{normal}(\gamma(Z^i_{(i)} \otimes D_i^{-2})\text{vec}(\widetilde{Y}^i_{(i)}), \gamma = [(Z^i_{(i)}(Z^i_{(i)})^T + \text{I}_{m_i}) \otimes D_i^{-2}]^{-1}).$$

B. Update D_i as follows.

(i) Sample $\{\text{vec}(Z^i)|\Lambda_i, D_i, \widetilde{Y}^i\}$ as in A(i).

(ii) Sample the elements of D_i^2 independently from

$$\{D_i^{-2}[j,j]|Z^i, Y, \boldsymbol{D_{-i}}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}\} \sim \text{gamma}\left((\nu_0 + m/m_i + k_i)/2, \text{rate} = \left(\nu_0 d_0^2 + J[j,j] + ||\Lambda_i[j,]||^2\right)/2\right)$$

where $J = (\widetilde{Y}^i_{(i)} - \Lambda_i Z^i)(\widetilde{Y}^i_{(i)} - \Lambda_i Z^i)^T$ and $||\cdot||^2$ denotes the Frobenius norm.

These updates for Λ_i and D_i are Metropolis-Hastings proposals with acceptance probabilities equal to one (see Appendix A).

Similar to step 1, the update of Σ_i in step 3 can be performed by sampling from the full conditional distribution of Σ_i^{-1} given the data and all other parameters:

$$\{\Sigma_i^{-1}|\boldsymbol{D}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma_{-i}}, Y\} \sim \text{Wishart}(\kappa_i + m/m_i, (\text{I}_{m_i} + \widetilde{Y}^i_{(i)}(\widetilde{Y}^i_{(i)})^T)^{-1}). \tag{13}$$

This proposal also has acceptance probability equal to one, and as in step 1 and 2, the latest updates of all other modes' parameters and the missing data are used in the calculation of $\widetilde{Y}^i$.

Step 4 of the algorithm that updates the missing data can be done in one update or as a sequence of updates for any partition of Y_m. All elements of Y_m can be updated together by sampling from the conditional distribution of $\text{vec}(Y_m)$ given the covariance parameters and observed data. Although the conditional distribution is normal, such an update can be expensive due to the large matrices

often involved in computing the distribution's covariance matrix. Updating the missing data in partitions of the array known as slices, where one mode index is fixed, avoids the need to work with such large matrices. For the mortality data, examples of slices include the data for country c for all ages, sexes, and years or the data at age a for all sexes, years, and countries. Hoff (2011) shows that for a separable covariance structure the conditional distribution for a slice of an array given the rest of the array can be written as array normal distribution. If the missing data in the slice is then conditioned on the observed data in the slice, a multivariate normal distribution is obtained that can be used to update the missing data. Calculating the conditional distribution of the missing elements in a slice of the array via this two-step conditioning procedure (once for the slice and once for the missing data within the slice) circumvents computation with unnecessarily large matrices. As in the update of Σ_i, the normal distributions mentioned here represent the full conditional distributions of Y_m and a subset of Y_m so the acceptance probabilities equal one. Note that although updating subsets of Y_m may be easier computationally, it has the potential to make the Markov chain less efficient and increase the number of samples needed to obtain an accurate approximation of the posterior distribution.

Unlike in the frequentist setting where divergence of the maximum likelihood estimation procedure indicates a lack of information in the data about the parameters, the posterior distribution of the parameters given the data will always exist. Although Bayesian parameter estimates are available, we should be aware of what information the estimates reflect. The posterior distribution of the parameters given the data is combination of the information in the prior distribution and the information in the data. Extreme similarity between the prior distribution and the posterior distribution suggests that little information is gained from the data and inference based on the posterior distribution is primarily a reflection of the information in the prior.

Hyperparameters

When there is little prior information about the parameters, it is common to choose hyperparameter values that result in diffuse prior distributions. We propose $\nu_0 = 3$ and $\kappa_i = m_i + 2$ for $\{i : k_i = m_i\}$ as default values for the SFA model. These values ensure that the first moments of the prior distributions are finite and represent some of the most diffuse distributions in the Wishart and

gamma families, respectively. They also have the following properties.

$$\mathrm{E}[\Sigma_i] = \mathrm{I}_{m_i} \qquad \mathrm{E}[D_i^2[j,j]] = 3d_0^2 \qquad \mathrm{E}[\mathrm{tr}(\Lambda_i \Lambda_i^T)] = 3k_i m_i d_0^2 \tag{14}$$

Prior information about specific mode covariance matrices may be limited, however an estimate $\widehat{\psi}$ of the total variance of the array, $\psi = \mathrm{tr}(\mathrm{Cov}[\mathrm{vec}(Y)]) = \prod_{i=1}^{K} \mathrm{tr}(\Sigma_i)$, may be available. This information can improve parameter estimation by centering the prior distribution of the total variance of the array around a reasonable value. Based on the expectations in (14) and the independence of the mode covariance matrices in the prior, the prior expected value of the total variance of the array will equal the estimate, $\mathrm{E}[\mathrm{tr}(\mathrm{Cov}[\mathrm{vec}(Y)])] = \widehat{\psi}$, if

$$d_0^2 = \widehat{\psi}^{1/R} \left[\left(\prod_{j:0<k_j<m_j} [k_j + 1] \right) \left(\prod_{i=1}^{K} m_i \right) 3^R \right]^{-1/R} \tag{15}$$

where $R = \sum_{i=1}^{K} \mathbb{1}\{0 \leq k_i < m_i\}$ is the number of modes with factor analytic covariance structure. In the event there is no prior knowledge about ψ and it is not of interest in the analysis, we propose taking an empirical Bayes approach and obtaining an estimate of it based on the data. Possible estimates include $\widehat{\psi} = \frac{m}{m_o}||Y_o||^2$ or $\widehat{\psi} = \frac{m}{m_o}||Y_o - \widehat{M}_o(X,\beta)||^2$ if the model has a non-zero mean, where m_o denotes the number of observed entries in Y. In the latter case, $\widehat{M}_o(X,\beta)$ represents an initial estimate of the mean for the observed data, such as the ordinary least squares estimate. Specification of d_0^2, ν_0, and $\{\kappa_i : k_i = m_i\}$ as described here weakly centers the prior distribution for the total variation in Y around the estimate $\widehat{\psi}$.

3.3 Testing for the mode ranks

It is often difficult to choose the number of factors for a single mode factor model. This problem is only more pronounced in the array case where the rank k_i must be specified for each mode. As is done in single mode factor analysis (Mardia et al. (1979)), a likelihood ratio test can be constructed to test between nested SFA models with ranks $(k_1, ..., k_K)$ and $(k_1^*, ..., k_K^*)$, where $k_i \leq k_i^*$ for all i. However, due to the large number of possible combinations of ranks, choosing the ranks using these likelihood ratio tests is challenging. Here we propose an alternative mode-by-mode rank selection procedure that suggests when the rank specified for a given mode is sufficient for capturing the dependence within that mode.

Suppose a K-way array Y is normally distributed, and define $\widetilde{Y}$ to be the array obtained when Y is standardized by its covariance matrix $\Sigma = \text{Cov}[\text{vec}(Y)]$: $\text{vec}(\widetilde{Y}) := \text{vec}(Y)\Sigma^{-1/2}$. The elements of $\widetilde{Y}$ represent independent standard normal random variables. Thus, to determine whether the covariance in mode i is captured by a proposed $(k_1, ..., k_K)$ SFA model, we can compute $\widetilde{Y}$ using the SFA mode covariance matrix estimates as in (1) and test whether the covariance matrix of the rows of $\widetilde{Y}_{(i)}$ equals the identity. The likelihood ratio test statistic for this test is

$$t = \frac{m}{m_i}[\text{tr}(\widehat{V}) - \log|\widehat{V}| - m_i], \tag{16}$$

where $\widehat{V} = \frac{m_i}{m}\widetilde{Y}_{(i)}\widetilde{Y}_{(i)}^T$, and has an asymptotic $\chi^2_{m_i(m_i+1)/2}$ distribution under the null hypothesis of an identity row covariance matrix. Note that rejecting this test suggests that a more complex covariance structure is needed for the mode. We propose the following rank selection procedure based on these mode specific tests.

0. Consider an SFA model with all $k_i = 0$. Obtain estimates of the covariance parameters using the maximum likelihood procedure in Section 3.1 and compute $\widetilde{Y}$ using the estimates.

1. For each mode i, define $R_i = \text{Cov}[\text{vec}(\widetilde{Y}_{(i)})]$ and test $H_0 : R_i = \text{I}_{m/m_i} \otimes \text{I}_{m_i}$ vs $H_1 : R_i = \text{I}_{m/m_i} \otimes V$, where V is an unstructured $m_i \times m_i$ covariance matrix, using a likelihood ratio test with test statistic given by (16).

2. If the test for mode i rejects and $\begin{cases} \delta(m_i, k_i + 1) > 0, \text{ increase the rank } k_i \text{ by one.} \\ \delta(m_i, k_i + 1) \leq 0, \text{ set the rank equal to } m_i. \end{cases}$

   If the test for mode i does not reject, fix k_i at its current value and perform no further tests on the mode. Obtain maximum likelihood estimates $\{\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Lambda}}, \widehat{\boldsymbol{D}}\}$ for an SFA model with the new ranks $(k_1, ..., k_K)$ and compute $\widetilde{Y}$ using these new estimates.

3. Repeat steps 1-2 until each mode has failed to reject a test.

The suggested ranks $(k_1, .., k_K)$ are those that result at the end of this procedure. Recall that $\delta(m, k) = \left[(m-k)^2 - (m+k)\right]/2$ represents the reduction in the number of parameters when using a k-factor analytic covariance matrix instead of an $m \times m$ unstructured covariance matrix. The rank increases in step 2 reflect that an unstructured covariance matrix is specified when a factor analytic covariance structure no longer provides a reduction in the number of covariance parameters.

The maximum number of SFA models that could be considered using this procedure is bounded by largest value of k_l such that $\delta(m_l, k_l) > 0$, where l denotes the array mode with the largest dimension m_l. To control the type I error rate of all mode tests to be α for an iteration of steps 1 and 2, the level of each mode test can be set to α^r where r is the number of modes being tested (i.e. the number that have rejected every test thus far). An example of this procedure is described in Section 4.2 for the mortality data.

4 Application to Human Mortality Database death rates

In this section we analyze death rates from the Human Mortality Database (HMD) (University of California, Berkeley and Max Planck Institute for Demographic Research, 2009) using an SFA model, compare our model to other covariance models, and obtain predictions for over four hundred missing death rates. We focus on death rates for 5-year time periods for populations corresponding to combinations of sex, age, and country of residence. Specifically, we consider death rates from 1960 to 2005 for 40 countries, both sexes and twenty-three age groups, $\{0, 1-4, 5-9, 10-14, ..., 105+\}$. These data are represented in a 4-way array $Y = \{y_{ctsa}\}$ of dimension $(40 \times 9 \times 2 \times 23)$, where y_{ctsa} is the log death rate for country c, time period t, sex s and age group a. We will refer to a set of age-specific death rates for a combination of country, time period, and sex as a mortality curve.

We begin this section by introducing a flexible piecewise polynomial mean model and show the residuals from this mean model exhibit dependence within each mode: age, time period, country, and sex. Using the likelihood ratio testing procedure presented in Section 3.3, we select ranks for an SFA model. The resulting SFA model is compared to models with simpler covariance structures using out-of-sample cross validation and is used to impute multiple years of missing death rates for Chile and Taiwan.

4.1 Mean model selection

As discussed in the Introduction, existing methods for analyzing mortality data model the death rates for different countries, sexes, and/or time periods separately. Such an approach can be inefficient due to the strong similarities between mortality rates within the same country, time period, or sex. For this reason, we propose a new joint mean model for the HMD data that acknowledges the relationships between mortality rates that share levels of one or more of these

factors.

Figure 1 shows mortality curves defined by the twenty-three age-specific death rates for the United States and Sweden in four time periods. The large spikes at age zero represent infant mortality, and the humps around age twenty, which are especially evident in males, are attributed to teenage and young adult accident mortality. The overall shapes of the mortality curves for each sex are similar across countries and time periods, however Sweden has considerably lower mortality levels during childhood and young adulthood compared to the United States. This suggests that a mean model for the data should allow for different curves across countries and time periods, yet still take advantage of the similarity between death rates within the same country, age group, or sex.

Drawing from the mortality literature and viewing mortality rates as function of age, we propose the following piecewise polynomial (PP) mean model:

$$\mathrm{E}[y_{cysa}] = \begin{cases} \phi^0 & : a = 0 \\ \phi^1 + a\phi^{11} + a^2\phi^{12} & : 1 \leq a < 20 \\ \phi^2 + a\phi^{21} + a^2\phi^{22} + a^3\phi^{23} & : 20 \leq a \end{cases} \qquad \phi^i = \alpha_c^i + \beta_t^i + \gamma_s^i. \tag{17}$$

This model distinguishes between the infant, childhood, and adult stages of mortality by fitting each with a separate polynomial, whose coefficients are composed of additive effects for country, time period, and sex. The constant term at age zero is necessary to model the steep decline from infant mortality to child mortality that is not well represented by a low degree polynomial. Parameter estimates for this model based on the data array can be obtained by minimizing the ordinary the least squares (OLS) criterion $\sum_c \sum_t \sum_s \sum_a [y_{ctga} - \mathrm{E}[y_{ctga}]]^2$, and since the model is linear in its parameters, the OLS estimates can be solved for algebraically.

One of the most commonly used models in demography for age-specific mortality measures is the Heligman-Pollard (HP) model (Heligman and Pollard (1980)). This model also uses eight parameters to parameterize a mortality curve, however it is typically used to model each mortality curve individually and is nonlinear and non-convex in the parameters making estimation extremely difficult (Hartmann (1987), Congdon (1993)). When the HP model is fit separately to the 684 HMD mortality curves for the 38 countries missing no death rates using OLS, it requires over 5,400 parameters and under the assumption of independent, homoscedastic errors has a Bayesian Information Criterion (BIC) value of $-17,288$. However, when the PP model is fit jointly to the

same data, it contains 376 parameters and has a BIC of $-52,436$. Due to the relative parsimony of the PP model, its superior fit in terms of BIC, and its straightforward estimation, it was selected as the mean model.

4.2 Excess dependence and SFA rank selection

The piecewise polynomial model in (17) is extremely flexible. To investigate its fit to the HMD mortality rates, we focused on a subset of the original data that contains no missing observations, specifically the $(38 \times 9 \times 2 \times 23)$ array that does not have death rates for Chile or Taiwan. The OLS fit this data explains 99.5% of the variation in mortality rates (coefficient of determination, $R^2 = 0.995$). However, there is interest in whether excess correlation exists in the residuals since modeling it can improve both predictions of missing values and the efficiency of parameter estimates. Ordinary least squares estimates of the parameters in (17) are equivalent to maximum likelihood estimates assuming independent normal errors. To evaluate this latter assumption, we computed the empirical correlations between the mean model residuals for countries, time periods, and age groups by matricizing the residual array with respect to each mode and computing a sample correlation matrix for the mode.

As mentioned in the Introduction, the distributions of these correlations have substantially more large positive values than would be expected under the assumption of independent errors. For example, speaking specifically to the temporal dependence, the average correlation between adjacent time periods, those one time period apart, and those two periods apart is 0.79, 0.54, and 0.26, respectively. The first two principal components of each correlation matrix are shown in Figure 2. The horseshoe pattern in the time period principal components and the clustering of countries within the same region suggests temporal and geographic trends in the data are not captured by the mean (Diaconis et al. (2008)). This indicates that even though the mean model contains several country specific and time period specific parameters, similarities between the mortality curves of certain countries and time periods is not being accounted for. The mean model already contains over 370 parameters and it would likely be nontrivial to modify it to capture all of the dependence seen in the residuals. For this reason, we consider incorporating a covariance structure to model this excess dependence. An array normal separable covariance structure could be specified, however it would add over one thousand parameters to the model. Therefore, we instead consider an SFA

model for the data with the PP mean with the belief that some of the residual mode covariance matrices may be well approximated by a low rank factor analytic structure.

As outlined in Section 3.3, suggestions for the SFA ranks can be obtained from a repeated likelihood ratio testing procedure. For the mortality data, we consider (k_c, k_t, k_s, k_a) SFA models where the ranks correspond to the country, time period, sex, and age covariance matrices, respectively. The standardized residual array $\widetilde{Y}$ for a (k_c, k_t, k_s, k_a) SFA model is defined as $\text{vec}(\widetilde{Y}) = (\text{vec}(Y) - \text{vec}(\widehat{M}))(\widehat{\Sigma}_a^{-1/2} \otimes \widehat{\Sigma}_s^{-1/2} \otimes \widehat{\Sigma}_t^{-1/2} \otimes \widehat{\Sigma}_c^{-1/2})$, where $\widehat{M}$ represents the PP mean model maximum likelihood estimate and $\widehat{\Sigma}_i$ is the SFA mode i covariance matrix estimate. The results from the iterative testing procedure are shown in Table 1. The first step in this process is to consider a (0,0,0,0) SFA model where all covariance matrices are diagonal. The likelihood ratio test statistics for this model are shown in the first row of Table 1 and the corresponding 0.05 level critical values are shown the last row. Since the test for each mode rejects the null hypothesis of independent, variance one errors, the rank of each mode is increased by one in the subsequent model, except for that for the sex mode. A rank one factor analytic structure for a (2×2) covariance matrix has more parameters than an unstructured covariance matrix so the sex covariance matrix is unstructured in the next model. A box around a test statistic in the table indicates the mode failed to reject the test for the first time. Recall that when a mode's test does not reject, the rank for that mode is fixed and not increased in later models. The table shows where the sex, time period, country, and age ranks become fixed at two, four, nine, and ten, respectively. Observe that after a mode's rank is fixed, the test statistic for that mode stays below the critical value in all subsequent models. Although the mode tests are not independent of the covariance structures fit in the other modes, this consistency supports the suggested ranks.

4.3 Out-of-sample cross validation

We evaluate the SFA model by comparing its out-of-sample predictive performance with two simpler covariance models that share the same PP mean model. The three covariance models considered are the following:

M1: Independent and identically distributed (i.i.d.) model

M2: Time covariance model (0,9,0,0)

M3: SFA model (9,4,2,10)

Table 1: Iterative testing procedure for the SFA ranks. Each row represents an SFA model and each entry is the likelihood ratio test statistic based on (16). The 0.05 level critical value for each test is given in the last row. A box around a statistic indicates that the mode does not reject the test for the first time and the rank is fixed in subsequent models.

SFA ranks	Likelihood ratio test statistic			
(k_c, k_t, k_s, k_a)	Country	Time period	Sex	Age
(0,0,0,0)	21,852	14,482	702	27,883
(1,1,2,1)	9,526	5,853	$\boxed{0}$	14,451
(2,2,2,2)	4,425	1,722	0	6,374
(3,3,2,3)	2,776	716	0	3,762
(4,4,2,4)	1,946	$\boxed{17}$	0	2,422
(5,4,2,5)	1,556	14	0	1,833
(6,4,2,6)	1,287	10	0	1,340
(7,4,2,7)	1,040	8	0	967
(8,4,2,8)	892	5	0	540
(9,4,2,9)	$\boxed{762}$	8	0	363
(9,4,2,10)	737	8	0	$\boxed{257}$
$\chi^2_{.95}$ critical value	742	45	3	276

M1 corresponds to the conventional ordinary least squares (OLS) approach where all errors are assumed independent and identically distributed with a common variance parameter. Based on the temporal nature of the data, a natural first step to incorporating a covariance model is to consider an unstructured covariance matrix for time as in M2. In general, country mortality rates are relatively stable over time so if the observed mortality for a given country, year, and age deviates from the mean model in one year, it is likely the observations deviate in the same direction in neighboring years.

Fifty cross validations were performed by removing a random 25% of the array, estimating each of the three models on the remaining data, and computing the mean squared error (MSE) between the observed values and the predicted values for the withheld entries. The predicted values for M1 are those from the OLS PP mean estimate. For M2 and M3, the predictions are the posterior mean estimates of the missing values from the Bayesian estimation procedure described in Section 3.2.

A prior distribution for the parameters in the PP model is needed to perform simultaneous Bayesian estimation for the mean and covariance parameters. The prior on the vector of PP coefficients is a mean zero normal distribution with covariance matrix $m(X^TX)^{-1}$, where X is the design matrix for the PP model for vec(Y) and $m = \prod_{i=1}^{K} m_i$. This is a relatively uninformative prior as it is over 30 times more diffuse than the corresponding unit-information prior (Kass and Wasserman (1995)). The hyperparameters were specified as described in Section 3.2 where the mean estimate $\widehat{M_o}$ used in $\widehat{\psi}$ is the OLS estimate of the PP model. Since M2 has no modes with factor analytic structure, the prior on the time covariance matrix is

$$\Sigma_t^{-1} \sim \text{Wishart}\left(n_t = m_t + 2, \frac{m\widehat{\psi}}{m_t} I_{m_t}\right).$$

This specification is necessary to preserve the property that E[tr(Cov[vec(Y)])] = $\widehat{\psi}$ under the prior.

The results from the 50 cross-validations are shown in Table 2. The MSE for the SFA model was less than that of the time covariance model for each of the 50 cross-validation runs, and the MSE for the time covariance model was always less than that of the i.i.d. model. In terms of average MSE, both the time covariance model and the SFA model significantly improve upon the i.i.d. model, and the SFA model out performs the time covariance model by nearly a factor of two. This is evidence that even with the extremely flexible PP mean model, the SFA covariance structure still improves model fit as it is able to estimate the similarity between mortality rates across countries, time periods, age groups, and sexes, and use this information in its predictions.

Table 2: Average and standard deviation of the mean squared errors from 50 out-of-sample cross-validation experiments.

	M1 (i.i.d.)	M2 (Time Covariance)	M3 (SFA)
Average mean squared error	0.02996	0.00729	0.00385
Standard deviation of mean squared errors	0.00084	0.00049	0.00034

4.4 Prediction of missing data

The imputation of missing death rates is an important application of modeling mortality data as information is often incomplete for countries lacking accurate death registration data. We now

consider the original $(40 \times 9 \times 2 \times 23)$ array of mortality rates with observations for Chile and Taiwan. Seven time periods of mortality information are missing for Chile (1960-1995) and two time periods for Taiwan (1960-1970), combining for a total of 414 missing entries in the array. The maximum likelihood estimation algorithm in Section 3.1 cannot accommodate missing data so we are unable to reselect the SFA ranks using the testing procedure. However, this larger data array contains only two additional countries so the SFA ranks $(9, 4, 2, 10)$ selected for the reduced data are used here. Predictions for the missing death rates were based on samples from the Metropolis-Hastings procedure, for which the effective sample sizes for the Monte Carlo estimates of all missing values was greater than 500.

In the left column of Figure 3, posterior mean predicted death rates and 95% prediction intervals are shown for Chile in 1990 and Taiwan in 1965. To visualize the impact of the SFA covariance model on the predicted death rates, we investigate the difference between the SFA predicted values and the fitted values based on the PP mean model. The SFA predictions, $\hat{y}_p$, are conditional on the observed mortality rates for all other countries and time periods, while the mean model fitted values, $\hat{y}_m$, are based only on the estimate of the PP model. We call these differences, $\hat{y}_p - \hat{y}_m$, "predictive residuals" since they are based on predicted values instead of observed values. These differences illustrates the changes in the predicted values by using the estimated dependence between residuals within modes of the array and conditioning on the observed mortality rates. The empirical residuals based on the PP mean model, $y - \hat{y}_m$, were computed for the United States and Australia, the two countries most highly correlated with Chile (estimated correlations of around 0.40). These residuals were also computed for Japan and West Germany, the two countries most highly correlated with Taiwan (estimated correlations of around 0.13). The middle column of Figure 3 shows the predictive residuals for Chile and Taiwan and the empirical residuals for these select countries. The last column contains the empirical residuals in 1995 and 1970 when mortality information is available for all countries. Observe that the plots in the middle column and last column are similar, demonstrating an overall positive association for both sexes and all country pairs. This illustrates how the model uses the relationship between the empirical residuals of Chile and other countries to predict Chile's deviations from the mean model in years when Chile data is missing. The ability to draw information across multiple country, year, and sex residuals to impute missing values is a critical strength of the SFA model that is not shared by other mortality models

or simpler covariance structures.

The empirical residuals for Chile shown in the last column may not show as strong of an association with the United States and Australia as one would expect from a posterior mean correlation estimate of 0.4. However, recall that the estimate of the country correlations is based on all time periods, sexes, and ages. Although we show adjacent time periods in this plot, the correlation between the country residuals in the period adjacent to the missing time period and the correlations in time periods furthest away are weighted equally in the estimate of the country correlation, and hence weighted equally in the imputation of the missing data. For example, the correlation between Taiwan and Japan's empirical residuals in 2000 and that in 1970 influence Taiwan's imputations in 1965 equally. This property is a consequence of the separability of the SFA covariance matrix. A more complicated non-separable covariance model would be required for the correlations between countries, ages, and sexes to be differentially weighted in the imputation based on the proximity of the observed data to the missing data.

5 Discussion

In this article we introduced the separable factor analysis model for array-valued data. Unlike the array normal model where all mode covariance matrices are unstructured, SFA parameterizes mode covariance matrices by those with factor analytic structure. Using covariance matrices with reduced structure decreases the number of parameters in the model considerably and allows mode covariance matrices to be estimated using maximum likelihood methods for any array dimension. Including a covariance structure in a model for multiway data can drastically improve mean model parameter estimation and missing data predictions in situations where dependence exists within modes that is not captured by the mean model. In an out-of-sample cross validation study with a large set of mortality data, the SFA model was shown to have superior fit compared to models with simpler covariance structures, even in the presence of an extremely flexible mean model. The SFA model was also shown to estimate which countries have similar deviations from the mean model and was able to use this information to predict multiple years of missing death rates.

An alternative extension of factor analysis to arrays was considered that resembles the higher order singular value decomposition (see Appendix B). This model has a non-separable covariance structure and can be viewed a submodel of the single mode factor analysis model for $\text{vec}(Y)$. We

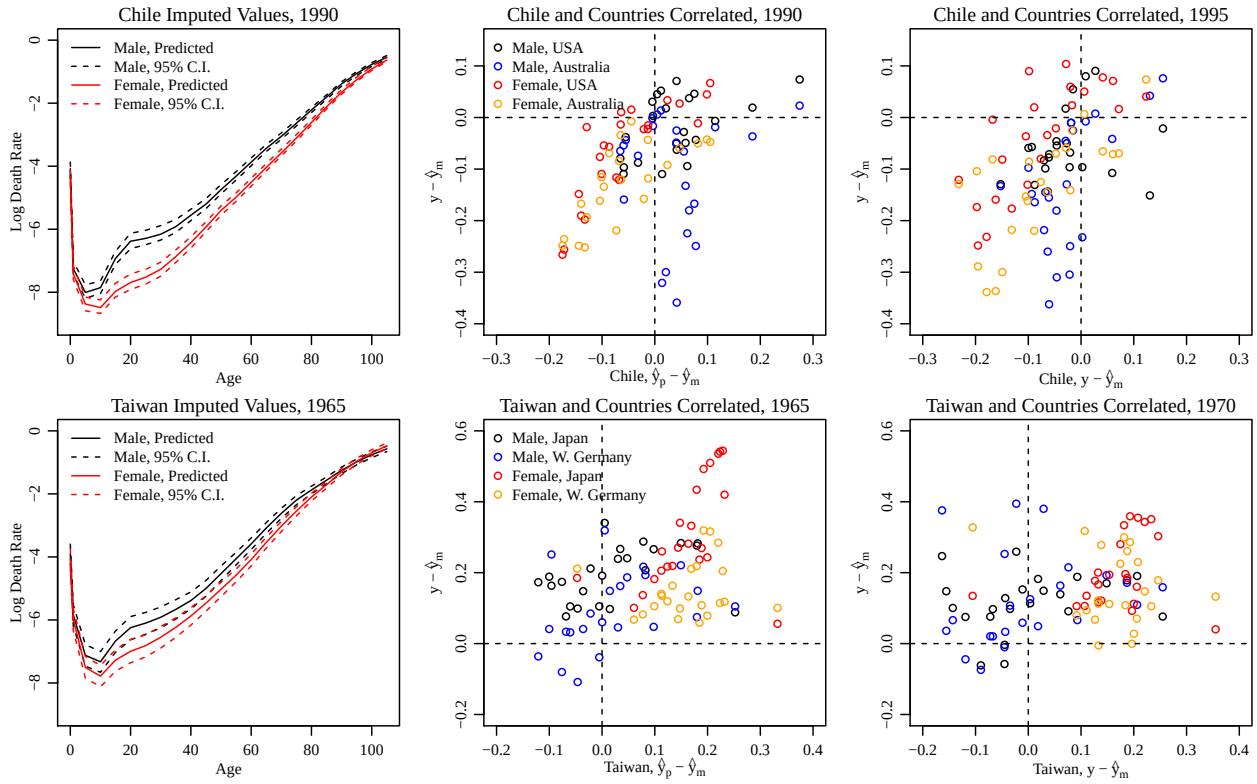

Figure 3: The first column of plots shows the predicted values and corresponding 95% prediction intervals for the missing death rates for Chile and Taiwan. The middle column shows the difference between the posterior mean predicted value and the piecewise polynomial mean function fitted value, $\hat{y}_p - \hat{y}_m$, for Chile and Taiwan, along with empirical mean model residuals, $y - \hat{y}_m$, for countries that are highly correlated with them in the posterior mean country covariance matrix. The last column contains empirical residuals for the following time period when Chile and Taiwan mortality is observed.

chose the SFA model over this alternative extension due to the interpretability of its parameters as single mode factor model parameters and for its use as an approximation to a separable covariance model with an unstructured covariance matrix in each mode.

A Appendix: Sampling Λ and D from their full conditional distributions

Let Λ_i^* be the proposed value of Λ that results from A(i-ii). The acceptance probability for this proposal is

$$\alpha(\Lambda_i^*, \Lambda_i) = \frac{p(\Lambda_i^*|Y, \Lambda_{-i}, D, \Sigma)p(\Lambda_i|\Lambda_i^*, D, \Sigma, \Lambda_{-i}, Y)}{p(\Lambda_i|Y, \Lambda_{-i}, D, \Sigma)p(\Lambda_i^*|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y)} = \frac{p(Y|\Lambda_i^*, \Lambda_{-i}, D, \Sigma)p(\Lambda_i^*|D_i)p(\Lambda_i|\Lambda_i^*, D, \Sigma, \Lambda_{-i}, Y)}{p(Y|\Lambda_i, \Lambda_{-i}, D, \Sigma)p(\Lambda_i|D_i)p(\Lambda_i^*|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y)}$$

The proposal probability can be written

$$\begin{aligned}
p(\Lambda_i^*|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y) &= \int p(\Lambda_i^*, Z^i|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y)dZ^i \\
&= \int p(\Lambda_i^*|Z^i, D, \Sigma, \Lambda_{-i}, Y)p(Z^i|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y)dZ^i \\
&= p(\Lambda_i^*|D, \Sigma, \Lambda_{-i}, Y) \int \frac{p(Z^i|\Lambda_i^*, D, \Sigma, \Lambda_{-i}, Y)}{p(Z^i|D, \Sigma, \Lambda_{-i}, Y)} p(Z^i|\Lambda_i, D, \Sigma, \Lambda_{-i}, Y)dZ^i \\
&= \frac{p(Y|\Lambda_i^*, D, \Sigma, \Lambda_{-i})p(\Lambda_i^*, D, \Sigma, \Lambda_{-i})}{p(D, \Sigma, \Lambda_{-i}, Y)} \cdot c(\Lambda_i, \Lambda_i^*|D, \Sigma, \Lambda_{-i}, Y) \\
&= \frac{p(Y|\Lambda_i^*, D, \Sigma, \Lambda_{-i})p(\Lambda_i^*|D_i)p(D)p(\Sigma)p(\Lambda_{-i}|D_{-i})}{p(D, \Sigma, \Lambda_{-i}, Y)} \cdot c(\Lambda_i, \Lambda_i^*|D, \Sigma, \Lambda_{-i}, Y)
\end{aligned}$$

where $c(\Lambda_i, \Lambda_i^*|D, \Sigma, \Lambda_{-i}, Y)$ represents the integral, which is symmetric in Λ_i and Λ_i^*. Plugging the last expression into the acceptance probability, we obtain $\alpha(\Lambda_i^*, \Lambda_i) = 1$. Analogous logic can be used to show the acceptance probability for a proposed D_i from B(i-ii) is also one.

B Appendix: Alternative extension of factor analysis to arrays

An alternative extension of factor analysis to arrays is motivated by the latent variable representation of single model factor analysis in (2). First consider extending the single mode factor model to estimate relationships among the rows and the columns of a matrix X, by writing

$$X_{p\times n} = \Lambda_1 Z \Lambda_2^T + D_1 E D_2^T = Z_{k_1\times k_2} \times \{\Lambda_1, \Lambda_2\} + E_{p\times n} \times \{D_1, D_2\}, \tag{18}$$

where $W \times \{A_1, ..., A_K\}$ indicates the first mode of the K-way array W is left multiplied by A_1, the second mode is left multiplied by A_2, etc. As in SFA, Λ_i is $(m_i \times k_i)$, and D_i is diagonal, however we only consider $k_i > 0$. If Z and E have the same distributional properties as in (2), the covariance matrix of X is $\text{Cov}[\text{vec}(X)] = (\Lambda_2\Lambda_2^T \otimes \Lambda_1\Lambda_1^T) + (D_2^2 \otimes D_1^2)$.

The analogous model for a K-way array Y has the following equivalent representations:

$$Y = Z \times \{\Lambda_1, ..., \Lambda_K\} + E \times \{D_1, ..., D_K\} \tag{19}$$

$$\text{Cov}[\text{vec}(Y)] = (\Lambda_K \Lambda_K^T \otimes ... \otimes \Lambda_1 \Lambda_1^T) + (D_K^2 \otimes ... \otimes D_1^2)$$

where Z is $(k_1 \times ... \times k_K)$, E is $(m_1 \times ... \times m_K)$, and these again have the same distributional properties as in (2). The second moment of a matricization of the array is written

$$\text{E}[Y_{(i)} Y_{(i)}^T] = \alpha_i \Lambda_i \Lambda_i^T + \gamma_i D_i^2 \qquad \alpha_i = \prod_{j \neq i} \text{tr}(\Lambda_j \Lambda_j^T) \qquad \gamma_i = \prod_{j \neq i} \text{tr}(D_j^2).$$

Observe that the second moment has k_i-factor analytic structure and will be unstructured if $\Lambda_i = D_i = \Sigma_i^{1/2}$, where $\Sigma_i^{1/2}$ is any non-singular $m_i \times m_i$ matrix.

The model in (19) has many similarities to the higher-order singular value decomposition (HOSVD) (Tucker (1966), De Lathauwer et al. (2000)). The HOSVD states that any K-way array Y can be written $Y = G \times \{U_1, ..., U_K\}$, where G is an all-orthogonal core matrix of the same dimension as Y and U_i is $(m_i \times m_i)$ satisfying $U_i^T U_i = I$ for $i \in \{1, ..., K\}$. The i^{th} slice of Y in the j^{th} mode is that which results by setting the j^{th} index of Y equal to i. An array is considered to be of reduced rank if slices of the core array G are zero. A K-way array of rank $(r_1, ..., r_K)$ can be expressed by the HOSVD with a core matrix G of dimension $(r_1 \times ... \times r_K)$ where each U_i is of dimension $(m_i \times r_i)$. The alternative array factor model in (19) can be written as $Y = M + \widetilde{E}$ where $M = Z \times \{...\}$ is the mean and common factor portion, and $\widetilde{E} = E \times \{...\}$ is the error component. The mean structure of this model resembles the HOSVD if Z is viewed as a core array. Although Z is not all-orthogonal and Λ_i and Σ_i do not satisfy $\Lambda_i^T \Lambda_i = (\Sigma_i^{1/2})^T \Sigma_i^{1/2} = I$, M can be rewritten using the HOSVD to obtain a new Z, Λ_i and Σ_i that satisfy the constraints. The error component $\widetilde{E}$ is then interpreted as accounting for variation in Y which is not captured by the reduced rank approximation M.

The covariance model in (19) is a submodel of the single mode $(\prod k_i)$-factor model for $\text{vec}(Y)$ since the covariance matrix is comprised of a reduced rank matrix plus a diagonal matrix. SFA is equivalent to this alternative extension if zero or one mode is specified with factor analytic structure. A drawback of the HOSVD and this alternative extension is that mode parameter values are difficult to interpret and cannot be considered independently of parameters in other modes. For this reason, we chose to focus on SFA as the primary extension of factor analysis to arrays.

References

Allen, G. I. and Tibshirani, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *Annals of Applied Statistics*, 4(2):764–790.

Brass, W. (1971). On the scale of mortality. *Biological aspects of demography*, pages 69–110.

Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology*, 37:1–21.

Carter, L. R. and Lee, R. D. (1992). Modeling and forecasting us sex differentials in mortality. *International Journal of Forecasting*, 8(3):393–411.

Chiou, J.-M. and Müller, H.-G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association*, 104(486):572–585.

Coale, A. and Demeny, P. (1966). *Regional Model Life Tables and Stable Populations*. Princeton University Press.

Congdon, P. (1993). Statistical graduation in local demographic analysis and projection. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 156(2):237–270.

Currie, I. D., Durban, M., and Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Diaconis, P., Goel, S., and Holmes, S. (2008). Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*, 2(3):777–807.

Felipe, A., Guillen, M., and Nielsen, J. P. (2001). Longevity studies based on kernel hazard estimation. *Insurance: Mathematics and Economics*, 28(2):191–204.

Genton, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics*, 18(7):681–695.

Hartmann, M. (1987). Past and recent attempts to model mortality at all ages. *Journal of Offiicial Statistics*, 3(1):19–36.

Heligman, L. and Pollard, J. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, 107(1):49–80.

Hoff, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6:179–196.

Human Mortality Database (2011). *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de.*

Jennrich, R. and Bobinson, S. (1969). A Newton-Raphson algorithm for maximum likelihood factor analysis. *Psychometrika*, 34(1):111–123.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32:443–482.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934.

Kiers, H. A. L. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14:105–122.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Kroonenberg, P. M. (2008). *Applied multiway data analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Lawley, D. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, 60(2):64–82.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting u.s. mortality. *Journal of the American Statistical Association*, 87(419):659–671.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the lee-carter method. *Demography*, 42:575–594.

Liu, C. and Rubin, D. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Stat. Sinica*, 8:729–747.

Manceur, A. M. and Dutilleul, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239(0):37–49.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Academic Press.

Martínez-Ruiz, F., Mateu, J., Montes, F., and Porcu, E. (2010). Mortality risk assessment through stationary space-time covariance functions. *Stochastic Environmental Research and Risk Assessment*, 24:519–526.

McNown, R. and Rogers, A. (1989). Forecasting mortality: A parameterized time series approach. *Demography*, 26(4):645–660.

Mode, C. and Busby, R. (1982). An eight-parameter model of human mortality - the single decrement case. *Bulletin of Mathematical Biology*, 44:647–659.

Murray, C. J. L., Ferguson, B. D., Lopez, A. D., Guillot, M., Salomon, J. A., and Ahmad, O. (2003). Modified logit life table system: Principles, empirical validation, and application. *Population Studies*, 57(2):165–182.

Oort, F. J. (1999). Stochastic three-mode models for mean and covariance structures. *British Journal of Mathematical and Statistical Psychology*, 52:243–272.

Renshaw, A. and Haberman, S. (2003a). Lee-carter mortality forecasting: a parallel generalized linear modelling approach for england and wales mortality projections. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):119–137.

Renshaw, A. and Haberman, S. (2003b). Lee-carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33(2):255–272.

Renshaw, A. and Haberman, S. (2003c). On the forecasting of mortality reduction factors. *Insurance: Mathematics and Economics*, 32(3):379–401.

Renshaw, A., Haberman, S., and Hatzopoulos, P. (1996). The modelling of recent mortality trends in united kingdom male assured lives. *British Actuarial Journal*, 2(02):449–477.

Robertson, D. and Symons, J. (2007). Maximum likelihood factor analysis with rank-deficient sample covariance matrices. *Journal of Multivariate Analysis*, 98:813–828.

Rubin, D. and Thayer, D. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47:69–76.

Siler, W. (1983). Parameters of mortality in human populations with widely varying life spans. *Statistics in Medicine*, 2(3):373–380.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292.

Stein, M. L. (2005). Spacetime covariance functions. *Journal of the American Statistical Association*, 100(469):310–321.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311.

United Nations (1982). Model life tables for developing countries. *Population Studies*, 77.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Zhao, J. H., Yu, P. L., and Jiang, Q. (2008). ML estimation for factor analysis: EM or non-EM? *Statistics and Computing*, 18(2):109–123.